Curtain call at the Cartesian theatre

Krzysztof Dolega & Joe Dewhurst*

The authors would like to thank Prof. Tobias Schlicht, Dr Dave Ward, Alessio Bucci, and two anonymous referees, for comments and helpful discussion.

Curtain Call at the Cartesian Theatre

Abstract
Hobson & Friston (2014) outline a synthesis of Hobson's work on dreaming and consciousness with Friston's work on the free energy principle and predictive coding. Whilst we are sympathetic with their claims about the function of dreaming and its relationship to consciousness, we argue that their endorsement of the Cartesian theatre metaphor is neither necessary nor desirable. Furthermore, if it were necessary then this endorsement would undermine their positive claims, as the Cartesian theatre metaphor is widely regarded as unsustainable. We demonstrate this point and then develop an alternative formulation of their position that does not require the Cartesian theatre metaphor. Our positive goal is to clarify Hobson & Friston's confusing usage of philosophical terminology, replacing it where possible with the more transparent language of the forward models framework. This will require some modifications to their account, which as it stands is philosophically and empirically unsustainable.

0 – Introduction

In 'Consciousness, Dreams, and Inference: The Cartesian Theatre Revisited', Hobson & Friston (2014) propose a new approach to understanding consciousness in the Bayesian brain framework. As the title of their article suggests, the authors attempt to reintroduce the Cartesian theatre as a metaphor in order to better understand the functional role of consciousness and dreaming in a hierarchically organized biological system governed by probabilistic inferences (the Bayesian brain). Whilst we support their project of unifying consciousness and dreaming explanations under the banner of probabilistic inference, we try to show that their appeal to the Cartesian metaphor does not yield any explanatory power that could help us understand the function of consciousness. Moreover, we argue that this theoretical commitment makes Hobson and Friston's position dangerously similar to Cartesian materialism, which is the view that there is a discrete neural locus of conscious experience (see section 2 for more details).

If our attack on the Cartesian view advocated by Hobson & Friston is successful, many of the authors' bold claims have to be abandoned. One of these is the claim that the Cartesian view can solve the hard problem of consciousness, along with the problems of free will and mental causation. For reasons of brevity we focus only on Hobson & Friston's claims about the causal power of qualia in relation to the Cartesian theatre, in order to illustrate that these claims lack argumentative support and may be inconsistent with the cognitive architecture presupposed by the authors. Nevertheless, since many of the proposals made by Hobson and Friston rely on their Cartesian assumptions, demonstrating that the view is incoherent will cast doubt on their other claims.

* Both authors contributed equally.

We will begin with an overview of the position put forward by Hobson & Friston (section 1). Having done that we will summarise Dennett's criticism of the Cartesian theatre metaphor and Cartesian materialism (section 2). Next we will discuss Hobson & Friston's possible commitment to Cartesian materialism, and explain why we think it is problematic (section 3). Finally we will propose an alternative formulation of their position that is not committed to Cartesian materialism, but preserves the authors' positive contributions (section 4). In concluding we will raise two further questions that arise when considering how Hobson & Friston's theory connects with other existing views on consciousness.

1 – Setting the stage: Hobson & Friston's proposal

Hobson & Friston's proposal about the nature of consciousness builds on the foundations of the recently developed action-oriented predictive coding framework (AOPC). Friston has played a central role in developing this framework (Friston 2008, 2010; Friston, Daunizeau & Kiebel 2009; Hohwy, Roepstorf & Friston 2008; Feldman & Friston, 2010), and has previously collaborated with Hobson on a related proposal (Hobson & Friston 2012). The AOPC framework assumes that the brain is performing hierarchically structured causal inference, where hypotheses about the possible causes of sensory input drive the system in a top-down manner. These hypotheses determine the behaviour of the lower levels of the system, cascading down the hierarchy to terminate with predictions about patterns of sensory receptor activations. The difference between the expected and the actual activity in the neural periphery, referred to as *prediction error*, is propagated upwards, causing model revision. Overall the system is driven to minimise error by discarding inaccurate or imprecise hypotheses, meaning that it will tend to settle on an accurate hypothesis that allows it to successfully navigate the world.

Before we move on to unpack Hobson & Friston's proposal about consciousness, we will illustrate how the predictive framework can explain the phenomenon of binocular rivalry (Bresse 1909). In this visual paradigm subjects are simultaneously presented with two different objects (eg. a house and a face), one in each visual hemifield. Surprisingly, most subjects do not report seeing the two objects as superimposed on one another, but rather experience steady switching between consciously perceiving one or the other. AOPC can easily accommodate this phenomenon (see Hohwy et al 2008). The brain creates two hypotheses (the 'face' and the 'house' hypothesis) that are equally probable due to there being the same amount of sensory evidence for both of them. However, because of previously obtained information about the basic principles about the causal structure of the world (a probabilistic prior), in this case, that two objects cannot occupy the same spatial location at the same time (and possibly due to priors for different categories of objects) the system resolves the conflict by introducing a temporal discrepancy between the two hypotheses, rather than merging them into one.

The example of binocular rivalry has become canonical in the AOPC literature as an illustration of the relationship between the framework's probabilistic architecture and subjective experience. The explanation of the unstable percept in terms of system's failure to settle on a singular hypothesis reveals one of the framework's central claims - that the contents of system's probabilistic hypotheses correspond to the contents of conscious experience. Although the exact nature of this relation is one

of the main issues in the debate between different proponents of the AOPC approach (see the discrepancy between positions endorsed by Clark [2012] and Hohwy [2014]), postulating such a relation is central to the endeavour of explaining our mental capacities in probabilistic terms.

Having provided an example that illustrates the basic assumption of the predictive framework, we can return to what Hobson & Friston have to say about consciousness:

> '[we] consider consciousness to be the process of perceptual inference about states of the world causing sensations […where] inference is taken to be the formation of *probabilistic beliefs* through optimizing the *sufficient statistics* of probability distributions. [… Consciousness is] finding the best (in a Bayes optimal sense) probabilistic explanation for our sensorium.'
> (Hobson & Friston 2014: 7)

Let's unpack that a little bit. States of the world cause sensations, resulting in a 'sensorium', which is then interpreted in order to form probabilistic beliefs that capture a statistically optimal prediction that best explains or predicts the sensory input. This process of perceptual inference is to be considered as constitutive of consciousness.

Hobson & Friston cast this proposal as 'inference based on the **private theatres of virtual reality**' (2014: 9, emphasis added), because the system's predictions can be said to be 'generated in a **virtual model of the world** and then tested against sensory reality' (*ibid*: 8, emphasis added). Here we can already see the emphasis that they place on the Cartesian theatre metaphor. Although the authors do not properly elucidate these terms, the ideas of an internal theatre and a virtual reality model of the world play a central role in their account of consciousness.

They find this interpretation of the AOPC framework especially attractive as it allows for a unified account of experience during sleep and wakefulness, compatible with their previously elaborated model of dreaming (see Hobson & Friston 2012). According to Hobson & Friston, the comparison between wakefulness and dreaming consists in the system's ability to self-generate 'fictive sensations', which are used to optimize the current model (in perception) or minimize complexity and redundancy (during sleep)(2014: 20). They add that this comparison can be made 'from both a phenomenological and neurobiological perspective' (*ibid*), presumably in the sense that dreaming not only feels like a simulacrum of waking experience, but also shares certain neural correlates associated with waking consciousness.

Despite appealing to 'private theatres […] that are so manifest in dreaming' (2014: 9), Hobson & Friston explicitly claim they do not want to invite Cartesian dualism about conscious experience. Their account is said to 'renounce dualism' (*ibid*: 7) in favour of a probabilistically understood dual-aspect monism[1] that 'provides a monistic solution that bridges the Cartesian divide between the *res*

1 Following Chalmers (1996) and Russell (1927) we take monism to be a position that assumes only one kind of fundamental substance. In this paper we understand materialistic monism to be the claim that all phenomenal or protophenomenal properties are in fact properties of physical matter. Any other variety of monism is held by us to claim 'certain phenomenal or protophenomenal properties as fundamental' (Chalmers 1996: 155) and non-reducible (*ibid*: 166).

cogitans and the *res extensa* […] where **immaterial beliefs** are probability distributions that are entailed by material sufficient statistics' (*ibid*: 7, emphasis added). The details of this proposal are not at all clear, but we believe that at the very least it amounts to a commitment to a 'virtual Cartesianism', where what Hobson & Friston call "immaterial beliefs" are understood as being embodied in probabilistically encoded neural states that become conscious when they are brought together for inferential processing. In section 3 we will argue that this proposal is unstable, collapsing either into an epiphenomenalism with regard to conscious beliefs, or requiring a stronger commitment to an unsustainable form of Cartesian materialism.

Although Hobson & Friston invoke Dennett when stressing that their account is non-dualistic (2014: 7), it is unclear whether their combination of the Cartesian theatre metaphor with dual-aspect theory is genuinely sustainable in light of Dennett's critique of Cartesian materialism. In the next two sections we will rehearse the problems with Cartesian materialism before examining whether or not Hobson & Friston's account can avoid them.

2 – A look behind the scenes: what is wrong with Cartesian materialism?

Dennett introduces the term 'Cartesian materialism' to refer to the view that there is a place in the brain where 'it all comes together' (1991: 107), some kind of discrete neural location where incoming sensory data is collated and becomes conscious.[2] Whilst the Cartesian materialist will be aware that there is no literal homunculus who sits and watches this show, this process of bringing together data remains metaphorically like a theatre because of the way in which the data is presented for conscious appreciation. While it is unclear whether Hobson & Friston have something like this in mind when they refer to the Cartesian theatre metaphor, their continual use of this metaphor and proclaimed denouncement of dualism suggest they might be endorsing a position dangerously similar to the one criticized by Dennett. We will first unpack the details of Dennett's criticism, before comparing it with Hobson & Friston's position in the following section.

Much of Dennett's work on consciousness has been aimed at opposing Cartesian materialism. According to him 'the brain is the headquarters, the place where the ultimate observer is, but there is no reason to believe that the brain itself has any deeper headquarters […] arrival at which is necessary or sufficient condition for conscious experience' (Dennett 1991: 106). Thus he thinks there can be no Cartesian theatre, metaphorical or otherwise. One of the main reasons that Cartesian materialism cannot work is that postulating a precise neural location for all conscious states would imply the possibility of determining the precise time at which a certain percept becomes conscious, by establishing when information from sensory transducers reaches the neural 'seat' of consciousness (i.e. the metaphorical Cartesian theatre). This, Dennett argues, has been shown to be impossible due to the findings of empirical paradigms that show significant temporal discrepancies between stimulation and verbal report, such as the colour Phi phenomenon (Kolers & von Grunau 1976) and Libet's experiments on action onsets and readiness potentials (Libet et al, 1983).

2 The claim that the moment of conscious experience is temporally synchronous (but neutrally distributed) can also be construed as a form of Cartesian materialism, but here we will focus primarily on the question of a discrete neural location. Much of what we have to say would apply equally to temporal synchronicity.

Moreover Dennett argues that postulating such a location is redundant, since the production and control of behaviour is carried out by the brain as a whole, the result of a multitude of parallel processes taking place simultaneously over many different time scales. To imply, for example, that visual processing is carried out in the occipital lobe only for the resulting data to then be passed on to the area of the brain that is 'really' in charge would invite a version of the homunculus fallacy (see Kenny 1984: 125). Our understanding of any cognitive function, consciousness and dreaming included, must involve an account of interactions with other processes taking place throughout the brain, rather than singling out any single discrete location that is totally responsible for a given function – or so says Dennett.

Dennett's position is very controversial, especially when it comes to consciousness, and has come under fire from both philosophers and cognitive scientists (see e.g. Chalmers 1995, Bogen 1992, Block 1995). One of the strongest replies to his arguments against Cartesian materialism has come in the form of an accusation of picking a fight with a straw man. O'Brian & Opie, though committed to attacking Denntt's project from a different angle, offer a compelling summary of this common criticism:

> 'Cartesian materialism, it is alleged, is an impossibly naïve account of phenomenal consciousness held by no one currently working in cognitive science or the philosophy of mind. Consequently, whatever the effectiveness of Dennett's demolition job, it is fundamentally misdirected' (O'Brien & Opie 1999: 941)

Whilst this accusation may have rung true in the past, Hobson & Friston do in fact explicitly embrace at least the Cartesian theatre metaphor, and as we will argue this means that they are also committed to a form of Cartesian materialism. At first glance this might seem to obviously be the case, as they denounce dualism whilst endorsing the existence of 'private theatres of consciousness', which seem to be the two most important criteria for Cartesian materialism. However it is somewhat unclear both whether they are genuinely materialists, as they openly claim that consciousness is an 'immaterial process' (Hobson & Friston 2014:8), and also whether they are genuinely committed to a discrete physical locus of consciousness, as required by Cartesian materialism. We will unpack their position, which they call 'dual-aspect monism', in the next section.

3 – 'Dual-aspect monism': starring Cartesian materialism

The rationale behind Hobson & Friston's adoption of dual-aspect monism stems from the authors' conviction that their position is founded on 'a dualism that distinguishes between the (conscious) process of inference and the (material) process that entails inference' (2014: 6). They further clarify that this is a distinction between probabilistic beliefs and the physical brain states that encode those beliefs (*ibid*). As we have mentioned before, the authors clearly state that they do not support substance dualism, and indeed, their position seems to be some kind of non-reductive property dualism (the identity between property-dualism and dual aspect theory has been asserted and clarified in a recent paper [Hobson, Hong and Friston 2014: 12]). They frequently make clear their commitment to the view that 'mental states (such as

probabilistic beliefs or qualia) are not ontologically reducible to physical states (such as neural states or sufficient statistics)'[3] (Hobson & Friston 2014: 21). However, their support for the non-reducibility of mental states and properties motivates the authors to label these states as 'immaterial' (*ibid*: 7), something that is ostensibly at odds with their purported physicalism[4].

The position expressed in 'Consciousness, Dreams and Inference' in fact seems closest to some version of a property dualistic token theory, where the relation between particular token physical states and corresponding token mental states is systematic and nomological, yet the token mental states cannot be said to be identical with the token physical states. This reading is plausible, as the authors' inferential approach to consciousness appears to require that particular mental states correspond to particular neural states in a unique and 'instantaneous' manner, without being strictly identified with those neural states (Hobson & Friston 2014: 21). The temporal co-occurrence of mental states with neural states is taken by us to mean that both kinds of states simultaneously bring about intentional action, resulting in a firm commitment to the causal efficacy of conscious mental states and properties: 'qualia are not just entailed (or induced) by physical states […] they determine the paths of those states' (*ibid* 26).

Despite these bold claims about the causal efficacy of mental states and properties, Hobson & Friston still profess to hold a physicalist position. Neither author claims that mental states or properties could be instantiated (or be causally efficient) without the specific neural substrates upon which they depend – to do otherwise would entail a *de facto* commitment to dualism. In this sense Hobson & Friston seem committed to some form of supervenience relation between the mental and the physical.[5] Nonetheless, the continual emphasis that they place upon the irreducibility and immateriality of conscious mental states, as well as their insistence on using the 'dual-aspect' moniker rather than simply acknowledging the supervenience relation, leaves them open to the accusation that they are committed to exactly the kind of residual Cartesianism that Dennett warns against.

Whilst most of the Cartesian terminology and imagery that Hobson & Friston deploy seems to serve a merely rhetorical purpose, they do openly endorse the idea of a Cartesian theatre: 'we are forced to consider something like a theatre when we discuss consciousness, especially when we consider the presence of a self or agent as

3 It is possible that, in labelling probabilistic beliefs as ontologically different from states of the physical world, the authors are treating probabilistic beliefs as mathematical entities. However, this does not explain why they are inclined to treat all mental properties as immaterial.

4 Although there are many varieties of physicalism we take physicalists to be committed to the main claims that mental states and properties are not fundamentally different or independent from physical states. Although they can take the relationship between the mental and physical to be reductive or non-reductive (Chalmers 1996: 166), a physicalist cannot claim that mental states are immaterial, since this would stand in direct contradiction to the idea that such properties are fundamentally physical. See Goff (2014) and Ney (2014) for a detailed discussion of the distinctions and differences between monism and physicalism.

5 In a more recent paper, Hobson, Hong and Friston (2014: 12) have claimed that their view is a kind of functionalism compatible with property dualism (what they call 'dual-aspect theory) as well as reductive and non-reductive varieties of physicalism. However, we find this 'clarification' difficult to understand in light of the present paper, as these three positions cannot all be equally compatible with the solution to the problem of mental causation offered by Hobson & Friston (2014) (see Kim [2005] for a full overview of the differences between property dualism and varieties of physicalism in respect to mental causation).

an integral part of the virtual reality model' (2014: 27). Even this endorsement, when taken together with their professed physicalism, might not be enough to commit Hobson & Friston to a Dennettian 'finishing line'. After all, it appears that they could fall back on their use of the term 'virtual reality' (Hobson & Friston 2014: 8-10) in order to claim that rather than having a physical locus in the brain the mental theatre that they describe simply has a virtual existence (i.e., is a virtual machine running on probabilistic hardware – whatever this might mean in this case). This could be a clever move on the part of Hobson & Friston, allowing them to continue to make use of Cartesian terminology whilst evading the threat of Cartesian materialism. However, we do not feel that it is a move that they are legitimately able to make.

Hobson & Friston are committed to a very particular computational architecture in which their talk of virtual reality is firmly anchored: 'the brain maintains a model or virtual reality that it uses to explain sensory inputs' (2014: 11). This virtual reality model of the external world is stored in the system's priors, which are used to construct probabilistically encoded beliefs (or hypotheses) about the proximal cause of incoming sensory input. Recall our earlier example of binocular rivalry and the role that prior information about the regularities obtaining in the world plays in hypothesis selection. In calling this model a theatre (even if only metaphorically) while being committed to some kind of nomological identity relation between the mental states and physical states of the system, Hobson & Friston seem to be endorsing an isomorphism between conscious experience and the physical states carrying information within the AOPC architecture. For example, the authors seem to think that certain properties of conscious states, e.g. qualia, are instantiated in virtue of corresponding physically realized probabilistic states.

Although, as we have noted, Hobson and Friston are unclear about their metaphysical commitments, we are forced to interpret their position in the above way in order to make sense of their claims about mental causation. If our interpretation is right, this correspondence would mean that whenever a most probable hypothesis (i.e. the one with the highest posterior probability) is selected for conscious presentation, this selection happens at a particular mental location realized at a particular neural location. Exactly what 'selection for conscious presentation' means is unclear on the AOPC framework (eg. Hohwy (2012) also fails to clarify this), but the authors' appeal to a notion of a theatre strongly suggests that this 'presentation' takes place at some determinate location. The authors' subscription to the idea that particular mental states are probabilistic states realized by particular neural mechanisms implies that this location is also physical.

This becomes problematic when we consider what happens when the AOPC architecture responds to an external stimulus. Any unexpected change in the environment will provoke an influx of prediction error, which the system has two ways of responding to. It can engage in model revision (or passive inference) by updating its priors and producing a new hypothesis that more closely matches the incoming signal, or it can act on the environment in order to make it a better fit for the current hypothesis (active inference), which it does by generating predictions of proprioceptive input that are then matched by appropriate bodily motion. In neither case is there any point at which an appeal to a Cartesian theatre would prove explanatorily relevant, as it is hard to see where, or in what sense, a hypothesis, model, or prediction could be 'presented' to an internal theatre, virtual or not.

What Hobson & Friston ignore in their treatment of the AOPC architecture as a theatre or virtual reality model is that it is difficult to delineate a clear boundary between the model itself (the priors) and the multiple hypotheses that are generated from that model. The winning hypothesis is generated via a process involving the recruitment and modification of the information available in a set of priors, responding to prediction error elicited from the comparison of the previously most probable hypothesis with incoming sensory data. What is relevant is that at any time, multiple hypotheses are generated, maintained and compared on different levels of the system. It is more accurate to speak of hypothesis generation and weighting taking place co-occurently, as the probability for all hypotheses currently available at a particular level is determined by a continuous distribution. Thus, an addition of a new hypothesis or an increase in the probability of an already present hypothesis (eg. 'the object is red') decreases the probability of other hypotheses on the same level (eg. 'the object is green' and 'the object is blue' both become less probable). Moreover, higher order 'winning' hypotheses may be consistent with more than one lower level hypothesis (this is supposedly what happens in the case of binocular rivalry), in which case the probability distribution over lower level hypotheses become more dependent on attentional modulation and bottom-up information.

In light of this it is difficult to make sense of the claim that a hypothesis is presented against the backdrop of prior expectations in any Cartesian sense. This would require a fundamental separation between hypotheses and expectations, but expectations play a crucial part in forming the hypotheses themselves. The distinction between the model and the predictions it generates is problematic, as past hypotheses can influence learning and the formation of new priors, which will be used recursively to generate future predictions.[6] Whilst models and hypotheses are distinct, they do not come apart as easily as Hobson & Friston seem to want them to, and so there can be no discrete theatre that is stable and independent from the AOPC architecture as a whole.

Given that on this picture the model can be treated as fulfilling the traditional roles of both percept (via sensory predictions) and motor commands (via proprioceptive predictions), it is very hard to see what explanatory value is gained from Hobson & Friston's appeal to the Cartesian metaphor. Action and perception are both the result of hypotheses generated within the model itself, not anything external to it. Moreover, it is unclear where the finishing line for conscious presentation could be located, as the processes of model revision and hypothesis generation take place simultaneously across multiple levels of the hierarchy (Clark 2013: 189-190).[7] Whilst hypotheses are compared with priors, this is a distributed and sub-personal (i.e. non-conscious) process. Talk of conscious presentation at this level of analysis is misguided, ruling out any literal interpretation of the Cartesian metaphor.

Hobson & Friston might move away from a literal interpretation of the Cartesian theatre when they postulate that the theatre of consciousness is a *virtual* model. For example, it is tempting to think that in waking consciousness the subject is presented with a conscious percept constructed by her perceptual system (eg. a

6 See Hohwy 2013; Clark 2013, forthcoming; for a full account of the architecture that we take Hobson & Friston to be committed to.

7 Hohwy (2013) discusses the problem of conscious content in the Bayesian brain without appealing to Cartesian terminology.

percept of a bottle of beer on the table). The subject can then decide to ignore or interact with this perceptual object. On this picture, the percept is generated for the subject to experience. This is, of course, a very naïve way of speaking, that should be treated metaphorically. However, one has to understand that the generative model includes not only a representation of the external world, but also a representation of the active agent coupled with that world. If a single predictive mechanism is responsible for both perception and action then there is no space to fit in even a metaphorical homunculus who pulls the levers and chooses action, let alone a whole virtual theatre. The system has to predict the effect that its actions will have on incoming sensory information and take this into account when picking the most accurate hypothesis – agency has to be unified with perception and action (Clark 2013: 194-195). If Hobson & Friston are seriously postulating that the theatre of consciousness should be interpreted as a virtual entity then they seem to imply that the agent who sits in this theatre must also be virtual. This renders their commitment to the role of conscious states (or properties like qualia) in causing behaviour (2014: 21) implausible. In VR, the system has to render not only the constructed environment but also the agent embedded into it. To predict how the environment can change with interactions, the VR model has to simulate the source of changes – the body, what it does, how it looks and changes over time. Since the whole brain is supposed to be governed by probabilistic inference, the sense of self and agency must also be products of this mechanism. On the AOPC the model of the body, including proprioception and interoception, will be crucial for the emergence of selfhood and a sense of agency. However, such model can exist only as one of the resources used by the wider system to navigate the world, not a separate or privileged 'driver' in the seat of consciousness. If Hobson & Friston's proposal is to interpret the subjective self as a separate model that interacts with the generative model, used to predict the world (the VR model) and how it changes with bodily interactions then it is difficult to see how this virtual agent could have genuine, rather than virtual or metaphorical, causal power. This virtual self might have a sense of agency, but no real causal efficacy, as the system's behaviour is fully determined by the AOPC mechanism – the actions performed by the body are fully determined by the probabilistic process of prediction error minimization

By invoking the Cartesian metaphor Hobson & Friston end up locked between a commitment to Cartesian materialism, which is not supported by the AOPC architecture, or a fully virtual Cartesian theatre, which lacks causal efficacy and explanatory power. They do not seem to recognise that this latter option is a philosophically naïve view which contradicts their own account of mental causation. Either conscious phenomena are causally efficacious in virtue of being realized at a particular physical locus, in which case Hobson & Friston end up being committed to Cartesian materialism (since they speak of this locus as a theatre), or the locus of consciousness is identified with a virtual construct, whose role and relationship to the wider system remains unexplained, rendering it epiphenomenal. This is the crux of our argument: Hobson & Friston's proposal is caught between the two horns of Cartesian materialism and epiphenomenalism. Moreover, we take this failure at navigating the problem of mental causation to bedilemma to be a direct result of unclear and misguided their use of obscure philosophical terminology, ultimately resulting in a failure to elucidate how consciousness can be accommodated by the AOPC framework.

4 – Putting an end to Cartesian theatrics: the Bayesian brain without the theatre

Our criticism of Hobson & Friston's approach is not meant to rule out the possibility of accounting for consciousness within a Bayesian framework. We find the framework eminently plausible, but want to resist the idea that it in any way supports a Cartesian metaphor. In the following section we will try to show that the advantages of Hobson & Friston's proposal can be preserved even after discarding the authors' Cartesian terminology and substituting it with concepts popular in simulation and emulation frameworks (Grush 1997, 2004; Pickering & Clark 2014).

One of the main advantages of the Bayesian framework is the prospect of obtaining a unified explanation of mechanisms responsible for perceptual and perception-like experiences in waking consciousness and dreaming (Clark: 2012). This is explored in Hobson's previous work on the AIM (activation, input–output gating, and modulation) model of proto-consciousness and dreaming (2009), which he elaborates on in previous papers with Friston (Hobson & Friston 2012, 2014).

The core principle behind this combined research project is the idea that 'dream consciousness and its physiological underpinnings [should be] considered as a virtual reality model of the world that prepares us for waking consciousness (...)' (Hobson & Friston 2014: 8). This proposal is aimed at explaining, in terms of free-energy minimisation, the biological function of dreaming and the evolutionary puzzle of the seeming suspension of homeothermy in REM sleep (Hobson & Friston 2014: 10). For Hobson & Friston, 'sleep is a necessary process that requires the (nightly) suspension of sensory input — so that synaptic plasticity and homoeostasis can reduce the redundancy and complexity accrued during wakefulness (Gilestro, Tononi and Cirelli, 2009).[8] In short, it is necessary to gate sensory input (and responses) to finesse the complexity of virtual reality models used to navigate the waking sensorium.' (Hobson & Friston 2014: 10)

The incorporation of the well established AIM model into the Bayesian brain framework is an interesting proposal, as it helps to flesh out one of the framework's central claims – that perception, hallucination, and dreaming are different working modes of the same prediction minimisation mechanism. Given these assumptions Hobson & Friston's proposal that dream consciousness (as well as wakeful experience and hallucination) can be metaphorically understood as a VR model of the world seems to be useful in elucidating the relationship between subjective experience and the system's structure. However, as we argued in the previous sections, the adoption of this metaphor does not entail a Cartesian view on consciousness.

One familiar way of making sense of Hobson and Friston's proposal regarding dreaming and consciousness is to understand the processes occurring during sleep as an emulation (or simulation) of the environment (cf. Metzinger 2003: 140; Revonsuo 1995, 2006). The emulation and simulation frameworks postulate that 'the brain constructs neural circuits that act as models of the body and environment' (Grush 2004: 377). The brain is said to implement forward models, running parallel to (emulation), or integrated into motor commands

8 Although see Tononi & Cirelli 2014 for evidence that this process of synaptic pruning might take place during NREM, rather than REM, sleep. This distinction is worth noting, but it does not impact greatly on anything that we have to say here.

(simulation) that, through modelling of the interactions between the body and environment, create feedback on the outcomes of actions before their performance is completed. As Grush elucidates in response to comments on his seminal 2004 paper, emulator (as well as simulator - see Pickering & Clark 2014) systems can not only run 'in parallel with the represented domain in order to form expectations that can be of use in sensory processing' (Grush 2004: 425), but also can be used 'off-line in order to see what a certain course of action might lead to (planning), or to train the controller (imagined rehearsal to improve skills), or just for fun (dreaming)' (Grush 2004: 425). While Grush may seem to dismiss dreaming as mere entertainment, the emulation/simulation frameworks can easily accommodate a more serious hypothesis about understanding processes occurring during REM sleep as the off-line use of models in processes of rehearsal and optimisation, consistent with Hobson & Friston's treatment of dream experiences as VR. In terms of AOPC, this is manifested as a process of decreasing the complexity of the predictive models (cf. Hobson, Hong, & Friston 2014; Hinton et al 1995).

The forward model frameworks can also facilitate the understanding of waking consciousness as a virtual model of the environment (reality). The AOPC architecture endorsed by Hobson & Friston is based on the idea of generative models - models that, through inference about the causes of regularities in the input, can create predictions about the future states of the perceptual systems. 'A generative model thus generates consequences from the causes in the same way that a forward model maps from causes to consequences. Forward models are thus examples of generative models.' (Pickering & Clark 2014: 1)

Applying the language of forward modelling to the VR metaphor helps to capture the framework's basic assumption about the relation between the model's predictions and phenomenal experience. As Pickering & Clark (2014) point out, the AOPC architecture integrates the forward model of perceptual expectations and motor commands. This means that whilst it is possible to decouple the system from the motor plant-body (as in the case of dreaming), the actions of the organism during wakefulness are mostly determined by its expectations about the contingencies obtaining between the external facts and the states of sensory organs, rather than by direct access to the environment. Through construction, deployment and continuous updating of the models of the environment the brain constantly tries to attune itself to the external world. This fine-tuning takes the form of honing the probabilistic representations that capture the structure of the perceptual and motor contingencies.

There is a serious debate about the extent to which this processing strategy can accurately produce information about the external facts. While some authors have put forward a claim that the probabilistic representations employed in the Bayesian brain are 'world revealing' (Clark 2012, 2013), this position has recently come under serious attack. Hohwy (2014) has pointed out that the Bayesian architecture poses a significant possibility for misrepresenting the environment, and that due to its mathematical grounding it assumes that no single hypothesis can be

tested to the point of certainty (in probability theory, the posterior probability of a hypothesis can never take a value equal to 1, which means that no amount of evidence and prior experience can result in absolute certainty). While Hohwy's arguments are aimed at the views about the embodied and extended nature of the mind, they also carry weight for the discussion of consciousness.

The strength of Hobson & Friston's consciousness-as-VR metaphor is that it can bring together insights from Hohwy's criticism of the quasi-direct account of perception championed by Clark with important facts about the integrated-forward-model nature of the AOPC architecture. It is widely assumed that on the probabilistic framework, the content and phenomenal quality of perceptual experiences is determined by the winning hypotheses/model (Hohwy, Roepstorf, & Friston 2008; Hohwy 2012, 2013; Clark 2012, 2013). Bringing this assumption together with the insight that even winning hypotheses have a degree of uncertainty that cannot be removed, and that these hypotheses are driving our perception and action as expectations about future sensory states, it seems that the content of our experiences must be something like a VR model of the world - a brain's construct of what is outside of itself.

5 – Curtain call at the Cartesian theatre

The aim of this paper was to show that Hobson and Friston's proposal to revive the Cartesian metaphor to explain consciousness and dreaming in the Bayesian brain is incoherent, and that the positive claims of their view can be preserved without the help of Cartesian terminology. While it is important to note that we do not think that Hobson and Friston's proposal is the one and only hypothesis about consciousness worth integrating into the prediction error minimisation framework, the explanation of the biological function of dreaming that is on offer is a valuable addition to other proposals about the role of consciousness.

Two important questions remain open regarding consciousness and the AOPC framework: how exactly are winning hypotheses linked to phenomenal experience, and what is the functional significance of a process being conscious in the Bayesian brain? Hobson and Friston's proposal does not offer any significant insight into how to address these very difficult problems.

Summing up, although the view and arguments used by Hobson and Friston are not free from deficiencies, we think that both the VR metaphor introduced in their paper, as well as the project of integrating the AIM model of dreaming into the Bayesian brain framework, are useful contributions to the field. We hope that in presenting an alternative way of thinking about Hobson & Friston's proposal, their ideas can be better integrated into the emerging framework.

References

Breese, B.B. (1909) Binocular rivalry, *Psychological Review*, **16** (6), pp. 410–5.

Block, N. (1995) On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, **18** (2), pp. 227-287.

Bogen, J.E. (1992). "Descartes' fundamental mistake: Introspective singularity". *Behavioral and Brain Sciences* (15): 184–247. Commentary on Daniel C. Dennett and Marcel Kinsbourne (1992) *Time and the observer: The where and when of consciousness in the brain.*

Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.

Chalmers, D.J. (1995) Explaining Consciousness: The 'Hard Problem', *Journal of Consciousness Studies*, **2** (3), pp. 200-219.

Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science, *Behav. Brain Sci.*, **36** (3), pp. 181–204.

Clark, A. (2012) Dreaming the Whole Cat: Generative Models, Predictive Processing, and the Enactivist Conception of Perceptual Experience, *Mind*, **121** (483), pp. 753–771.

Dennett, D.C. (1991) *Consciousness Explained*, London: The Penguin Press.

Feldman, H., Friston, K.J. (2010) Attention, Uncertainty, and Free-Energy, *Frontiers in Human Neuroscience*, **4** (215), [Online], http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3001758/ [24 Oct 2014].

Friston, K. (2010) The free-energy principle: A unified brain theory?, *National Review of Neuroscience*, **11** (2), pp. 127-138.

Friston, K. (2008) Hierarchical Models in the Brain, *PLoS Computational Biology*, **4** (11), [Online], http://www.fil.ion.ucl.ac.uk/~karl/Hierarchical%20Models%20in%20the%20Brain.pdf [4 Oct 2014].

Friston, K.J., Daunizeau, J., Kiebel S.J. (2009) Reinforcement Learning or Active Inference?, *PLoS ONE*, **4** (7), [Online], http://www.fil.ion.ucl.ac.uk/~karl/Reinforcement%20Learning%20or%20Active%20Inference.pdf [14 July 2014].

Goff, P. (2014) Against Constitutive Russellian Monism, in Alter, T., Nagasawa, Y. (eds.), *Russellian Monism*, Oxford: Oxford University Press, [Online], http://www.philipgoffphilosophy.com/uploads/1/4/4/4/14443634/against_constitutive_russellian_monism.pdf [12 November 2014].

Gilestro, G.F., Tononi, G., Cirelli, C. (2009) Widespread changes in synaptic markers as a function of sleep and wakefulness in Drosophilia, *Science*, **324**(65), pp. 109-112.

Grush, R. (2004) The emulation theory of representation: motor control, imagery, and perception, *Behavioral and Brain Sciciences*, **27** (3), pp. 377–442.

Grush, R. (1997) The architecture of representation, *Philosophical Psychology* **10** (1), pp. 5–25.

Hinton G.E., Dayan, P., Frey, B.J., Neal, R.M. (1995) The "wake-sleep algorithm for unsupervised neural networks, *Science*, **268** (5214), pp. 1158-1161.

Hobson, J.A., (2009) The AIM Model of Dreaming, Sleeping, and Waking Consciousness, in Squire, L.R. (ed.), *Encyclopedia of Neuroscience*, Oxford: Academic Press, pp. 963–970.

Hobson, J.A., Hong, C.-H., Friston, K.J. (2014) Virtual reality and consciousness inference in dreaming, *Frontiers in Psychology*, **5** [Online], http://journal.frontiersin.org/journal/10.3389/fpsyg.2014.01133/full [16 November 2014].

Hobson, J.A., Friston, K.J. (2014) Consciousness, Dreams, and Inference: The Cartesian Theatre Revisited, *Journal of Consciousness Studies*, **21** (1-2), pp. 6–32.

Hobson, J.A., Friston, K.J. (2012) Waking and dreaming consciousness: Neurobiological and functional considerations, *Progress in Neurobiology*, **98** (1), pp. 82–98.

Hohwy, J. (2014) The self-evidencing brain, *Noûs*, 48 (1) [Online], http://onlinelibrary.wiley.com/doi/10.1111/nous.12062/full [15 Sept 2014].

Hohwy, J. (2013) *The Predictive Mind*, Oxford: Oxford University Press.

Hohwy, J. (2012) Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3(96) [Online], http://philpapers.org/archive/HOHTSB.pdf [24 October 2013]

Hohwy, J., Roepstorff, A., Friston, K.J. (2008) Predictive coding explains binocular rivalry: an epistemological review, *Cognition*, **108** (3), pp. 687-701.

Kenny, A. (1984) The homunculus fallacy, In idem, *The legacy of Wittgenstein* (pp. 125–136). Oxford: Blackwell. (First published in Grene, M. (ed.), *Interpretations of life and mind*, London: Routledge & Kegan Paul, 1971).

Kiebel, S.J., Daunizeau, J., Friston, K.J. (2009) Perception and Hierarchical Dynamics. *Frontiers in Neuroinformatics*, **3** (20), [Online], http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718783/pdf/fninf-03-020.pdf [27 October 2014]

Kim, J. (2005) *Physicalism, or Something Near Enough*, Princeton: Princeton University Press.

Kolers, P., von Grunau, M. (1976). Shape and color in apparent motion. *Vision Research*, **16** (4), pp. 329-335.

Libet, B., Wright, E.W., Gleason, C.A. (1983) Readiness potentials preceding unrestricted spontaneous pre-planned voluntary acts. *Electroencephalographic and Clinical Neurophysiology*, **54** (3), pp. 322–325.

Metzinger, T. (2003) *Being No One. The Self-Model Theory of Subjectivity,* Cambridge: MIT Press.

Ney, A. (2014) A Physicalist Critique of Russellian Monism, in Alter, T., Nagasawa, Y. (eds.), *Russellian Monism*, Oxford: Oxford University Press, [Online], https://rochester.edu/college/faculty/alyssaney/pdfs/russellianmonism.pdf [12 November 2014].

O'Brien, G., Opie, J. (1999) A Defense of Cartesian Materialism, *Philosophy and Phenomenological Research*, **59** (4), pp. 939-63.

Pickering, M., Clark, A. (2014) Getting Ahead: Forward Models and their place in Cognitive Architecture, *Trends in Cognitive Sciences*, **18** (9) , pp. 451–456.

Revonsuo, A. (2006) *Inner Presence: Consciousness as a Biological Phenomenon*. Cambridge: MIT Press.

Revonsuo, A. (1995) Consciousness, dreams and virtual realities. *Philosophical Psychology,* **8**(1), pp. 35-58.

Russell, B. (1927) *The Analysis of Matter*, London: Kegan Paul.

Tononi, G. & Cirelli, C. (2014) Sleep and the price of plasticity, *Neuron,* **81**(1), pp. 12-34